Americas Conference on Information Systems (AMCIS)

AMCIS 2004 Proceedings

Association for Information Systems *Year* 2004

KBSVM: KMeans-based SVM for Business Intelligence

Jiaqi Wang
University of Technology, Sydney

Chengqi Zhang
University of Technology, Sydney

KBSVM: KMeans-based SVM for Business Intelligence

Jiaqi Wang
Faculty of Information Technology,
University of Technology, Sydney
Australia
jqwang@it.uts.edu.au

Chengqi Zhang
Faculty of Information Technology,
University of Technology, Sydney
Australia
chengqi@it.uts.edu.au

ABSTRACT

The goal of business intelligence (BI) is to make decisions based on accurate and succinct information from massive amounts of data. Support vector machine (SVM) has been applied to build the classification model in the field of BI and data mining. To achieve the original goal of BI and speed up the response of real-time systems, the complexity of SVM models should be reduced when it is applied into the practical business problems. The complexity of SVM models depends on the number of input variables and support vectors. While some researchers have tried to select parts of input variables to build the model, this paper proposes a new method called KMeans-based SVM (KBSVM) to reduce the number of support vectors. The experiments on real-world data show that the KBSVM method can build much more succinct model without any significant degradation of the classification accuracy.

Keywords

BI, SVM, KMeans, cluster, KBSVM.

INTRODUCTION

The goal of business intelligence (BI) is to make well-informed decisions based on accurate and succinct information extracted from massive amounts of practical data. It is very difficult for decision makers to manually analyze massive amounts of data. Valuable information hidden behind data can be mined with the help of many technologies from different areas such as data warehouse, data mining, web services, etc. These technologies together compose the architecture of BI.

Data mining in BI is to automatically extract the patterns and build the models for the practical business problems. Building the classification model is one of the most important issues in BI and data mining, i.e. detecting which customers of telecommunication companies always pay later than allowed. Many methods have been explored to build the classification model such as Decision Tree, Neural Networks. These methods have been embedded into some data mining products, i.e. "IBM Intelligence Miner".

Since 1990's, the support vector machine (SVM) method has been paid more attention by data mining researchers and practicers because many experiments on real-world data show the SVM method can get the higher classification accuracy compared with other classical methods [1, 5]. Some researchers have applied it into the field of BI, i.e. the classification problem of direct marketing [6]. Recently the SVM method has been embedded into the data mining product "Oracle Data Mining Release 10g".

While the SVM method can have the high classification accuracy, the complexity of SVM models should be reduced for some reasons when it is applied into the practical business problems. One reason is that the original goal of BI is to summarize massive amounts of data into succinct information as possible. The other reason is that succinct models can reduce the response time of some real-time monitoring systems, i.e. financial market surveillance and network intrusion detection.

The complexity of SVM models depends on two factors: the number of support vectors and input variables. Although [6] has tried to reduce the complexity of SVM models by selecting parts of input variables, the complexity of SVM models still may be high because too many support vectors may be contained in the model, e.g. several thousands of support vectors in the experiment 1. This paper proposes a new method called KMeans-based SVM (KBSVM) to find fewer but necessary support vectors. In detail, the KBSVM method firstly clusters the original data and then uses the traditional SVM method to build the model based on the new clustering centers.

KMeans-based clustering has the advantages of summarizing data and preserving the data distribution. In this paper, the experiments on real-world data show that the KBSVM method can build much more succinct model compared with the traditional SVM method without any significant degradation of the classification accuracy. In the real data mining task (the experiment 1), the traditional SVM method builds the model including about 6000 support vectors while the KBSVM method can build the model including only 100 support vectors without any significant degradation of the classification accuracy.

This result means that the model built by the KBSVM method is almost 60 times faster than that built by the traditional SVM method. So the KBSVM method is more attractive than the traditional SVM method when it is applied into some real-time monitoring systems. Furthermore, we discover that for almost all seven "Adult" data sets in the experiment 1, about 100 support vectors just can make the classification accuracy up to 80% (close to that of the traditional SVM method). This implies that about 100 support vectors should be enough for this data mining task about "Adult".

The remainder of this paper is organized as follows. The next section introduces the KBSVM method and the third section gives some experiments to verify the effectiveness of KBSVM. The last section draws a conclusion about the KBSVM method.

KMEANS-BASED SVM

The theory about SVM is originally established by Vapnik et al [5] and this method has been applied to solve many practical problems since 1990's. SVM benefits from two good ideas: "maximizing the margin" and "kernel trick". These good ideas can guarantee the high classification accuracy of models and overcome the problem about the curse of dimension.

For the classification problem, the SVM method solves the quadratic optimization problem in (1).

$$min\ ||w||^2 + C \sum_{i=1...l} \xi_i\ ,\ s.t.\ (<w, x_i> + b)\ y_i \geq 1 - \xi_i\ ,\ \xi_i \geq 0,\ i=1...l \qquad (1)$$

This optimization problem is described geometrically in Figure 1. In addition, the kernel functions are used in the SVM method to solve the non-linear classification problems. Now the popular kernel functions include the polynomial function, the Gaussian radius basis function (RBF), and the sigmoid function. An example of solving the non-linear classification problem is described in Figure 2. The SVM model is represented as follow $f(x) = \sum_{i=1...n} \alpha_i\ K(x, x_i) + b$, where n is the number of support vectors, x_i is the support vector, $K(\bullet)$ is the kernel function and b is the bias.

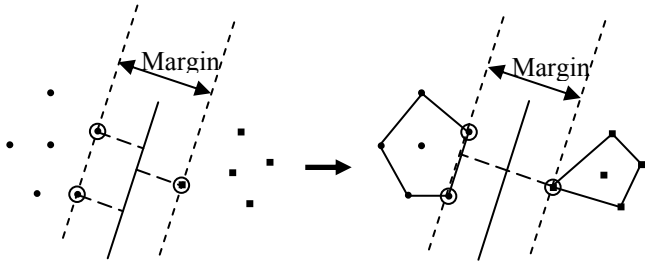

Figure 1. (from [7]) SVM maximizes the margin between two linearly separable sample sets. Maximizing the margin is equivalent to finding the shortest distance between two disjoint convex hulls spanned by the two linearly separable sample sets.

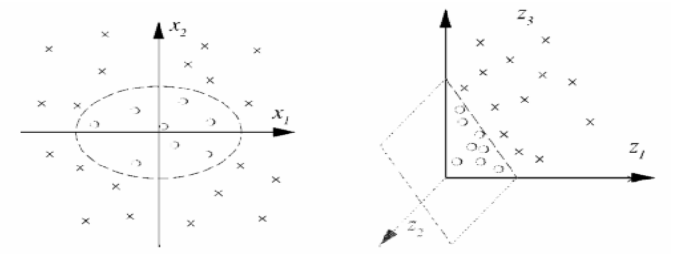

Figure 2. (from [4]) The samples are mapped from the 2-dimensional original space to the 3-dimensional feature space. The non-linear classification problem is converted to the linear classification problem by using feature mapping and kernel functions.

The complexity of SVM models should be reduced when it is applied into the practical business problems. On one hand, the original goal of BI is to summarize massive amounts of data into succinct information as possible. On the other hand, it is valuable to reduce the complexity of the SVM model when applied into some real-time monitoring systems, i.e. financial market surveillance and network intrusion detection.

The complexity of SVM models mainly depends on two factors: the number of input variables and support vectors. While [6] reduces the complexity of SVM models by selecting parts of input variables, the complexity of SVM models still may be high because too many support vectors may be contained in the model, e.g. several thousands of support vectors in the experiment 1. This paper tries the KMeans clustering technique to reduce the number of support vectors such that the SVM model can become more succinct without any significant degradation of the classification accuracy.

KMeans is a classical clustering method in the field of machine learning and pattern recognition [2]. It can provide the good summarization of data and preserve the data distribution. These advantages imply that it is possible to build the model with

the acceptable classification accuracy based on data compressed by KMeans clustering. In another word, there may be a lot of redundant information in the original data set.

An example about KMeans clustering is shown as Figure 3. There are almost twenty positive and negative samples in this example and they are reduced by KMeans clustering to only six cluster centers *C1, C2, ... , C6*. Moreover, the statistical distribution of the original data does not drastically degrade. Therefore the KBSVM method can build the model with fewer support vectors compared with the traditional SVM method and the classification accuracy can be guaranteed to some extent. The details of KBSVM are described in Figure 4.

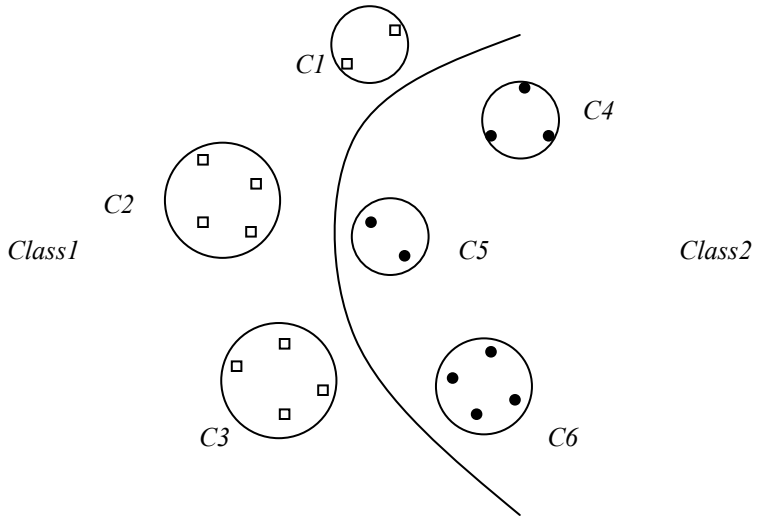

Figure 3. KMeans Clustering on 2-Dimensional Data Set. For the classification between *Class1* and *Class2*, *H* is the separating hyper-plane and *C1, C2, … , C6* are six clusters for *Class1* and *Class2* respectively.

Step1, set the acceptable compression rate according to users' requirement;
Compression Rate = Number of Original Data / Number of Cluster Centers
Step2, cluster each class of data respectively using the KMeans clustering technique;
Step3, generate the clustering centers as the new data trained;
Step4, train the compressed set using the traditional SVM method;
Step5, build the KBSVM model.

Figure 4. Details of KBSVM

EXPERIMENTS ON REAL-WORLD DATA SETS

This section provides some experiments on two real-world data set to verity the effectiveness of KBSVM. The definition of *"Compression Rate"* in Tables 1-8 can be found in Figure 4. Gaussian RBF kernel is used in all the following experiments. All the following experiments are performed on a computer with Pentium4 CPU and 256M Memory. We use the software "*LIBSVM 2.4*" [9] to implement the traditional SVM method. The results show that it is possible to obtain a much more succinct model without any significant degradation of the classification accuracy.

Experiment 1

The goal of this experiment is to predict whether the household has an income greater than $50,000 using the census form of a household. It is required to build the both accurate and succinct classification model. Data about this data mining task is from UCI "Adult" benchmark data set. We use seven sets "Adult 1-7" with training and testing data to compare the KBSVM method with the traditional SVM method from two sides: "the classification accuracy" and "the number of support vectors". The values of all parameters in this experiment are set as follows: Gaussian RBF kernel *sigma = 10* and the penalty factor in (1) *C = 10*. The results obtained by the KBSVM method are shown in Tables 1-7.

These tables show that the bigger compression rate (that is, the fewer cluster centers), the less the number of support vectors. Moreover, the support vectors usually can be compressed 7-8 times without any significant degradation of the classification

accuracy. This phenomenon is most obvious for the experiment on "Adult-7". When the support vectors are reduced more than 50 times (5861/117), the classification accuracy of KBSVM is almost the same as that of SVM (84.81%/83.63%). Furthermore, we discover that for almost all 7 "Adult" data sets, about 100 support vectors just can make the classification accuracy up to 80% (close to that of the traditional SVM method). This implies that about 100 support vectors should be enough for this data mining task. Obviously, the model including 100 support vectors can lead to the faster response of some real-time monitoring systems compared with the traditional SVM method.

Compression Rate	0	10	20	30	40	50	60	70	80	90	100
Number of SVs	**648**	**85**	44	33	24	23	18	15	13	13	10
Test Correct Rate (%)	**84.33**	**82.09**	76.46	76.23	75.95	75.95	75.95	75.95	75.95	75.95	75.95

Table 1. The Number of Support Vectors and Test Correct Rate of the Model by SVM and KBSVM on "Adult-1" Data Set

Compression Rate	0	10	20	30	40	50	60	70	80	90	100
Number of SVs	**947**	**121**	62	45	36	29	23	20	20	16	12
Test Correct Rate (%)	**84.55**	**82.09**	76.41	76.15	76.01	76.01	76.01	76.01	76.01	76.01	76.01

Table 2. The Number of Support Vectors and Test Correct Rate of the Model by SVM and KBSVM on "Adult-2" Data Set

Compression Rate	0	10	20	30	40	50	60	70	80	90	100
Number of SVs	**1233**	**157**	**85**	59	49	38	30	26	25	21	18
Test Correct Rate (%)	**84.52**	**82.59**	**78.03**	76.44	75.97	75.94	75.94	75.94	75.94	75.94	75.94

Table 3. The Number of Support Vectors and Test Correct Rate of the Model by SVM and KBSVM on "Adult-3" Data Set

Compression Rate	0	10	20	30	40	50	60	70	80	90	100
Number of SVs	**1808**	220	**123**	**84**	63	52	44	37	32	30	25
Test Correct Rate (%)	**84.49**	83.97	**82.82**	**80.91**	79.82	78.77	76.33	76.11	76.05	76.05	76.38

Table 4. The Number of Support Vectors and Test Correct Rate of the Model by SVM and KBSVM on "Adult-4" Data Set

Compression Rate	0	10	20	30	40	50	60	70	80	90	100
Number of SVs	**2380**	283	157	**110**	**86**	68	56	49	42	39	35
Test Correct Rate (%)	**84.44**	84.02	83.52	**83.13**	**82.9**	81.68	79.34	77.52	77.26	76.77	76.23

Table 5. The Number of Support Vectors and Test Correct Rate of the Model by SVM and KBSVM on "Adult-5" Data Set

Compression Rate	0	10	20	30	40	50	60	70	80	90	100
Number of SVs	**4094**	479	258	183	145	**116**	**97**	84	74	68	59
Test Correct Rate (%)	**84.59**	84.23	83.43	83.27	83.53	**83.24**	**82.02**	82.16	79.05	80.65	79.53

Table 6. The Number of Support Vectors and Test Correct Rate of the Model by SVM and KBSVM on "Adult-6" Data Set

Compression Rate	0	10	20	30	40	50	60	70	80	90	100
Number of SVs	**5861**	662	355	239	190	157	137	**117**	**103**	**92**	88
Test Correct Rate (%)	**84.81**	84.69	84.27	83.91	84.15	82.56	83.68	**83.63**	**83.39**	**82.78**	82.47

Table 7. The Number of Support Vectors and Test Correct Rate of the Model by SVM and KBSVM on "Adult-7" Data Set

Experiment 2

This experiment is performed on U.S. Postal Service (USPS) data [3], which is often used to evaluate the performance of the classification model based on the SVM method. It includes a lot of real handwritten digits (1-10) similar to those shown in Figure 5. There are 7291 training samples and 2007 testing samples in USPS data. The values of all parameters in this experiment are set as follows: Gaussian RBF kernel *sigma = 8* and the penalty factor in (1) *C = 5*. These results in Table 8

show that we can use much fewer support vectors to separate ten hand-written digits without any significant degradation of the classification accuracy. So this experiment can also verify the effectiveness of KBSVM.

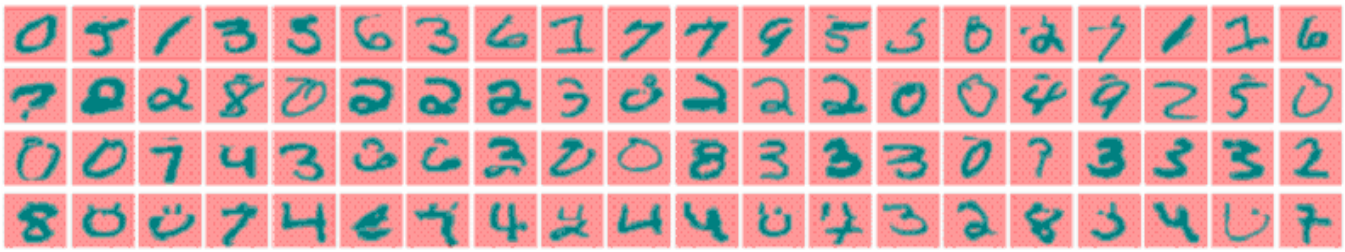

Figure 5. (from [8]) Normal and Atypical Hand-Written Digit

Compression Rate	0	10	20	30	40	50	60	70	80	90	100
Number of SVs	**1450**	324	**190**	137	126	99	83	70	65	60	54
Test Correct Rate (%)	**95.47**	93.42	**92.53**	91.88	91.98	91.38	90.88	90.23	90.33	90.33	89.64

Table 8. The Number of Support Vectors and Test Correct Rate of the Model by SVM and KBSVM on USPS Data Set

CONCLUSION

On one hand, decision makers wish to get the succinct and accurate model (and/or information). On the other hand, succinct models can reduce the response time of some real-time monitoring systems. The complexity of SVM models depends on the number of support vectors and input variables. This paper proposes the KBSVM method to reduce the number of support vectors. The experiments on real-world data show the effectiveness of this method. In the future, the KBSVM method should be tested on more practical problems of BI such as direct marketing, financial market surveillance, network intrusion detection, etc.

REFERENCES

1. Boser, B. E., Guyon, I. M. and Vapnik, V. (1992) A Training Algorithm for Optimal Margin Classifiers, *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory,* ACM Press.
2. Duda, R. O. and Hart, P. E. (1972) Pattern Classification and Scene Analysis, Wiley, New York.
3. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. J. (1990) Handwritten Digit Recognition with Back-propagation Network, *Advances in Neural Information Processing Systems*.
4. Muller, K. R., Mika, S., Ratsch, G., Tsuda, K. and Schölkopf, B. (2001) An Introduction to Kernel-based Learning Algorithms, *IEEE Transactions on Neural Networks,* 12, 2, 181-201.
5. Vapnik, V. (1999) The Nature of Statistical Learning Theory, 2nd ed., Springer-Verlag, New York.
6. Viaene, S., Baesens, B., Van Gestel, T., Suykens, J. A. K., Van den Poel, D., Vanthienen, J., De Moor, B. and Dedene, G. (2001) Knowledge Discovery in a Direct Marketing Case using Least Squares Support Vector Machines, *International Journal of Intelligent Systems*, 16, 9, 1023-1036.
7. Wang, J. Q., Tao, Q. and Wang, J. (2002) Kernel Projection Algorithm for Large-scale SVM Problems, *Journal of Computer Science and Technology*, 17, 5, 556-564.
8. Wang, J. Q., Zhang, C. Q., Wu, X. D., Qi, H. W. and Wang, J. (2003) SVM-OD: a New SVM Algorithm for Outlier Detection, *Foundations and New Directions of Data Mining Workshop in IEEE International Conference of Data Mining*.
9. http://www.csie.ntu.edu.tw/~cjlin/libsvm.